PubH 7470: STATISTICS

FOR TRANSLATIONAL & CLINICAL RESEARCH

CLINICAL TRIALS:

“AE” MONITORING IN PHASE II TRIALS

Phase I and II clinical trials present special difficulties because they involve use of agents whose spectrum of **toxicity** and likelihood of **benefits** are poorly understood/defined.

"AE" is the common abbreviation for "**Adverse Effects**"; also referred to as "**Side Effects**". I'll use the terms interchangingly for these "**unwanted events**".

In phase II trials, "efficacy" is the "outcome of interest" whereas "safety" is embedded to serve as "stopping rule"

In planning a clinical trial of a new treatment, we should always be aware that severe, even fatal, **side effects are a real possibility**. If the accrual or treatment occurs over an extended period of time, we must anticipate the need for a decision to stop the trial – **at any time** - if there is an excess of these unwanted events.

FOCUS ON PHASE II TRIALS

- In phase I trials, toxicity may be considered the "Outcome Variable" and dose escalation plan serves as the stopping rule.
- In phases II trials, we start to focus on efficacy which requires conventional analysis at the end. "Response" becomes the Outcome Variable; however, toxicity (or other adverse effects) may still turn out to be a problem during the trial.
- The monitoring of side-effect events is a separate activity that may require special consideration

Two-stage designs stops trials for "Efficacy Reason"; here we want rules to stop trials for "Safety Reason" – both, not treated enough or excessive adverse effects, put patients at risk.

Two-stage Designs are optional (decision by investigators) but stopping rules for safety reason are "required" by regulatory affairs agencies/entities.

For practical use, the "rule" has to be simple. At larger institutions, statisticians usually have to monitor these events on a daily basis.

SEQUENTIAL PROCESS

- The most common method for monitoring toxicity or adverse effects is to design a formal sequential "stopping rule" based on the limit of acceptable side-effect rates; the sequential nature of the rule allows investigators to stop the trial as early as the evidence that the event's rate becomes excessive.
- In multi-site trials, a "data safety and monitoring board" (DSMB) is required; in local phase II trials, it's the statistician's responsibility to form the rule and the Clinical Trial Office's staff is responsible for its implementation.

“BONE MARROW” BASICS

- “Bone marrow” is a spongy tissue found inside the bones; it contains “stem cells” that produces the body’s blood cells including white blood cells which fight infection.
- In patients with leukemia (& others), the stem cells malfunction producing excessive defective cells which interfere with the production of normal white and red blood cells; the defective cells also accumulate in the blood stream and invade other tissues/organs.
- Bad bone marrow needs to be replaced: BMT

“BMT” BASICS

- In “bone marrow transplant” (BMT), the patient’s diseased marrow is destroyed (usually by radiation); the healthy marrow is then infused into the patient’s bloodstream.
- In successful BMT, the new bone marrow migrates to the cavities of the bones (i.e. engrafts), and begins producing normal blood cells.
- If the marrow from a donor is used, the transplant is called “ allogeneic BMT” or “syngeneic BMT” if the donor is identical twin. If the donor used is from the patient (after treated), the transplant is called “autologous BMT”- lower success rate.

BMT: THE RISKS

- Bone marrow transplantation (BMT) is a complex procedure that exposes the patients to high risk of a variety of complications, many of them associated with death; these risks are in exchange for even higher risks associated with the leukemia or other disease for the patient is being treated.
- Since the patients' immune systems are weakened or destroyed, a complication which usually develops in half or more of BMT patients is "graft-versus-host" disease (GVHD)

SEVERE SIDE EFFECTS

- One way to prevent GVHD is to treat the donor's marrow prior transplantation; unfortunately, such a treatment may cause some patients with "engraftment" problems (either delayed or failed).
- The patient's own marrow was destroyed in preparation for BMT, if the donor's marrow does not engraft, the patient does not have the capacity to produce blood cells - and the transplant failed.
- A sequential monitoring for "non-engraftment" is desirable so as not to have more failed transplants.

The most common method for monitoring toxicity or adverse effects is to design a formal sequential "stopping rule"; and a sequential stopping rule could be formed in two different ways:

(i) a Bayesian approach to evaluating the proportion of patients with side effects, or

(ii) a Hypothesis testing approach - using the sequential probability ratio test (SPRT) - to see if the normal, acceptable side- effect's rate has been exceeded.

HYPOTHESIS TESTING

- Let start with the hypothesis testing approach because it's more "conventional" (with statisticians)
- Let π be the proportion of patients with adverse side effects; the problem becomes testing for the null hypothesis H_0: $\pi = \pi_0$ against alternative H_A: $\pi = \pi_A$; where π_0 is the normal baseline side-effect's rate (say, 5%) and is the "maximum tolerated rate" (say, 20%) - anything over that are considered excessive.

STATISTICAL MODEL

- We can assume that the number of adverse events "e" follows the usual Binomial Distribution $B(n, \pi)$, where n is the total number of patients.
- This leads to the log likelihood function:
  $L(\pi;e) = \text{constant} + e \ln\pi + (n-e) \ln(1- \pi)$

SEQUENTIAL PROBABILITY RATIO TEST

- When "e" adverse events are observed out of n "evaluable" patients, the test for null hypothesis H_0: $\pi = \pi_0$ against alternative H_A: $\pi = \pi$ can be based on "the log likelihood ratio statistic" LR_n: $LR_n = e(\ln \pi_A - \ln \pi_0) + (n-e)[\ln (1-\pi_A) - \ln (1-\pi_0)]$
- In conventional sequential testing, the statistic is calculated as each patient's evaluation becomes available and plotted against n; the trial is stopped if the plot goes outside predefined boundaries which depends on pre-set type I and type II errors.

SEQUENTIAL STOPPING RULE

- In testing for null the hypothesis H_0: $\pi = \pi_0$ against the alternative H_A: $\pi = \pi_A$, the decision is:

(i) to stop the trial and reject H_0 if $LR_n \geq \ln(1-\beta)-\ln\alpha$

(ii) to stop the trial and accept H_0 if $LR_n \leq \ln\beta-\ln(1-\alpha)$

(iii) continue the study otherwise

- In (i) there are too many events and in (ii) there are too few events - enough to make a decision.

SIDE EFFECTS MONITORING

- We do not stop the trial because there are too few events; we only stop the trial early for an excess of side effects, that is when:
  $e(\ln\pi_A - \ln\pi_0)+(n-e)[\ln(1-\pi_A)- \ln(1-\pi_0)]\geq \ln(1-\beta)-\ln\alpha$
- The lower boundary is ignored; trial continues
- Solving equation for “e” yields for upper boundary
- We can also solve the same equation for n .

RESULT

Stop the trial as soon as n, as a function of e, satisfies the following equation:

$$n(e) = \frac{\ln(1-\beta) - \ln\alpha + e[\ln(1-\pi_A) - \ln(1-\pi_0) - \ln\pi_A + \ln\pi_0]}{\ln(1-\pi_A) - \ln(1-\pi_0)}$$

n(e) is the number of evaluable patients for having e of them with adverse effects.

Rule: To stop the trial when we have "e" adverse effects before reaching a total of "n(e)" patients.

EXAMPLE

- Consider a simple case where we know that the baseline rate is $\pi_0 = .03$ or 3% and investigator sets a ceiling rate of $\pi_A = .15$ or 15%.
- If we pre-set the level of significance at $\alpha = .05$ and plan to reach of statistical power of 80% ($\beta = .20$), the the trial should be stop as soon as: n(1)=-7.8, n(2)=5.4, n(3)=18.6, n(4)=31.8 etc… rounding off to {-, 5,18, 31,…}.
- The "-" sign indicates that the first event will not result in stopping; the trial is stopped if "2 of the first 5, 3 of 18, or 4 of 31 patients have side effects"

Example: With the rule "{-, 5,18, 31,…}", the trial is stop if "the 18th patient was the 3rd side-effect event"

WEAKNESSES

- The hypothesis testing-based approach has two problems/weaknesses:

(i) At times, the result might appear to be "over aggressive"; the trial is stopped when the "observed rate" of adverse events (i.e. p=e/n) is below the ceiling rate π_A.

(ii) The statistical power falls short of the pre-set level because we apply the rejection rule for a two-sided test to a one-sided alternative.

IS IT REALLY OVER AGGRESSIVE?

- Take the example where we know that the baseline rate is $\pi_0 = 3\%$ and investigator sets a ceiling rate of $\pi_A = 15\%$; the stopping rule is: {-, 5,18, 31,…}.
- But, at the 4th event, the observed rate is 4/31 or 12.9%, still below the ceiling set at 15%.
- In the context of the statistical test, at that point, even though the observed rate is only 12.9% but enough to reject H_0 (3%) and "accept" H_A(15%), a rate at which the trial should be stopped.

Still kind of unsettling to a clinician to stop trial when the observed rate is still not yet considered unsafe (to him/her). Actually, the rule {-, 5,18, 31,…} is not very aggressive. In addition, the problem only appears so when the clinician is "too aggressive" to "go on" by setting the ceiling rate ways over the baseline rate (15% versus 3%). It would not appear as a problem when the "gap" is set smaller; for example, if know that the baseline rate is $\pi_0 = 3\%$ and investigator sets a ceiling rate of $\pi_A = 10\%$; the stopping rule would be: {-,-,10,23}. Here, we did not stop before the ceiling rate.

ABOUT STATISTICAL POWER

- The problem with statistical power, that it falls short of the pre-set level because we apply the rejection rule for a two-sided test to a one-sided alternative, is real!
- We can compute the actual/achieved power and compare to the pre-set power.
- For example, we decide to enroll a total of N patients and came with the rule LR_N; the true power is 1-Pr(N; π_A ,LR_N) where Pr(N; π_A ,LR_N) is the probability of reach N patients without having stopped the trial.

EXAMPLE

Suppose the rule is $LR_N = \{-, n(2), n(3), N\}$ and let u, v, and w be the numbers of adverse events that occur in each of the three segments of the trial [0,n(2)], [n(2),n(3)], and [n(3),N]. The probabilities for the three segments are $b[u;n(2), \pi_A]$, $b[v;n(3)-n(2), \pi_A]$, and $b[w;N-n(3), \pi_A]$ where $b[i;n, \pi_A]$ is the binomial probability to have exactly "i" events in n trials when the true rate is π_A. Reaching N patients without stopping the trial means that u<2, v<3-u, and w<4-(u+v). The true power is:

$$1 - \Pr(N; \pi_A, LR_N) = 1 - \sum_{u=0}^{1} \sum_{v=0}^{2-u} \sum_{w=0}^{3-u-v} b[u; n(2), \pi_A] b[v; n(3) - n(2), \pi_A] b[w; N - n(3), \pi_A]$$

By a similar calculation, but replacing π_A by π_0, we can calculate and check for the "size" of the test (type I error rate). For example:

$$1 - \Pr(N; \pi_0, LR_N) = 1 - \sum_{u=0}^{1} \sum_{v=0}^{2-u} \sum_{w=0}^{3-u-v} b[u; n(2), \pi_0] b[v; n(3) - n(2), \pi_0] b[w; N - n(3), \pi_0]$$

SOLUTION?

- The problem of being under-powered is correctable; since the power falls short, the boundary needs to be pulled downward to retain the pre-set level.
- For example, with $\pi_0 = 3\%$ and $\pi_A = 15\%$; the stopping rule found for 80% power was: {-7, 5,18, 31,…}; the true power is only 74%; we need to stop - say - for the 4th event before n(4) = 31.
- But when? Or How?

SOLUTION

- Goldman (1987) described an algorithm for computing exact power (and type I error rate).
- Goldman and Hannan (2001) proposed to repeatedly use that algorithm to "search" for a stopping rule which almost achieve the pre-set levels of type I error rate and statistical power; they also provided a FORTRAN program allowing users to set their own size and power (and design parameters); called G&H algorithm.

ABOUT G&H ALGORITHM

- Goldman and Hannan's algorithm works but choosing one between many rules found sometimes is not an easy job; several found could be "odd"!
- The gain may be small; it is true that the power falls short without a correction, but it's only a few percentage points.
- It does not solve the perceived problem that the observed rate may be below the pre-set ceiling rate.
- May be it would be more simple just to set the power higher, say 85% when we want 80%.

THE BAYESIAN APPROACH

Consider a Binomial distribution B(n,π), if we assume that the probability π has a "prior" distribution say - Beta(α,β); after "e" adverse events having observed, the "posterior" distribution of π becomes Beta (α+e,β+n-e). From this:

$$P(\pi_*) = \Pr(\pi > \pi_*)$$

$$= 1 - \int_0^{\pi_*} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+e)\Gamma(\beta+n-e)} y^{e+\alpha-1}(1-y)^{n-e+\beta-1}\,dy$$

MEHTA AND CAIN'S RULE

- By assuming an "uniform prior" (where $\alpha=\beta=1$), Mehta and Cain (1984) provided a simple formula:

$$P(\pi_*) = \Pr(\pi > \pi_*)$$

$$= 1 - \int_0^{\pi_*} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+e)\Gamma(\beta+n-e)} y^{e+\alpha-1}(1-y)^{n-e+\beta-1}\,dy$$

$$= \sum_{i=0}^{e} b[i; n+1, \pi_*]$$

- and proposed a rule for which the trial is stop when $P(\pi_0)$ is large, say exceeding 97%, where π_0 is the baseline side-effect's rate.

EXAMPLE

$$.97 = \binom{n(1)+1}{0}\pi_0^0(1-\pi_0)^{n(1)+1} + \binom{n(1)+1}{1}\pi_0^1(1-\pi_0)^{n(1)}$$

$$.97 = (1-\pi_0)^{n(1)+1} + \{n(1)+1\}\pi_0(1-\pi_0)^{n(1)}$$

$n(1) \cong 8$ when $\pi_0 = .03$

EXAMPLE

$$.97 = \binom{n(2)+1}{0}\pi_0^0(1-\pi_0)^{n(2)+1} + \binom{n(2)+1}{1}\pi_0^1(1-\pi_0)^{n(2)} + \binom{n(2)}{2}\pi_0^2(1-\pi_0)^{n(2)-2}$$

$$.97 = (1-\pi_0)^{n(2)+1} + \{n(2)+1\}\pi_0(1-\pi_0)^{n(2)} + \frac{[n(2)+1]n(2)}{2}\pi_0^2(1-\pi_0)^{n(2)-1}$$

$n(2) \cong 21$ when $\pi_0 = .03$

By applying the Mehta and Cain's Bayesian rule, we come up with pairs of numbers [e,n(e)]; it works just as the stopping rule obtained from the hypothesis testing-based approach. The major difference is that this Bayesian rule does not require the setting of a 'ceiling rate". At first it appears reasonable: if the usual normal rate is π_0 then the trial should be stopped when this rate is exceeded because the rate is no longer "normal"

EXAMPLE

With $\pi_0 = .03$ or 3%, the Mehta and Cain's rule yields the stopping rule {8, 21, 38,…}; that is to stop at 1 event out of 8 patients, 2 out of 21, 3 out of 38, and so on. As a comparison, with $\pi_0 = 3\%$ and $\pi_A = 15\%$; the test-based stopping rule found for 80% power was: {-, 5,18, 31,…} - to stop at 2 events out of 5 patients, 3 out of 18, 4 out of 38, and so on.

Goldman (1987), after consulting her collaborators/clinicians, concluded that even though the Mehta and Cain's Bayesian boundaries are philosophically very attractive but rather liberal, especially that it allows for the stopping of a trial after a single event. In fact, it seems too aggressive to trial simply because $\pi > \pi_0$; say when $\pi_0 = 3\%$ and $\pi = 3.5\%$ because patients benefit from the treatment as well.

MODIFICATIONS?

To overcome having an over-aggressive Bayesian rule, Goldman (1987) considered to raised the cutpoint ".97" for the posterior probability or formulating rule using $P(\pi_A)$ - instead of $P(\pi_0)$ - where π_A is the ceiling or maximum tolerated rate. For example, "the trial is stop when $P(\pi_A)$ is large, say exceeding 95% or 97%". However, she concluded that "various adjustments did not seem to remedy the problem".

It is true that setting a stopping rule based on large values of $P(\pi_0)$, say when $\pi_0 = 3\%$ and $\pi = 3.5\%$, may be too aggressive; the increase in the rate may not be large enough to be clinically significant (or to outweigh the benefits of the treatment).

On the other hand, setting a stopping rule based on large values of $P(\pi_A)$ alone seems "unsettling" because it ignores the baseline rate and never reveals the impact of the treatment on having side effects. It is true that setting a ceiling rate is always "subjective"; but by seeing both - π_0 and π_A - one would know how reasonable the parameters are.

To have a fair comparison with the corresponding hypothesis-based stopping rule, may be we should stop the trial based on large values of $P(\pi_A)$, say "the trial is stop when $P(\pi_A)$ is large, say exceeding 80% or 90%" - whatever the number usually used as the pre-set value for statistical power- not 97%. But this would make the resulting Bayesian rule even more aggressive!

The problem was the choice of the ‘prior”. With the “uniform prior” (where $\alpha=\beta=1$), the mean is .5; we really need some prior distribution with an expected value more in line with the concept of “rare” side effects.

Usually, in Bayesian analysis, the choice of the prior carries only moderate weight - sometimes not that important, a non-informative prior does the job. But here we conduct most very small trials and using sequential rule, it carries very heavy weight. For example, if we observe 3 events from 7 patients then (i) the posterior mean is still .5 (4/8) (leaning to stopping) with choice $\alpha=\beta=1$, but (ii) the posterior mean is .1 (4/40) (leaning to non-stopping) with choice $\alpha=1$ and $\beta=32$

OPTIONS

- There are no perfect choice for a prior
- Uniform prior may be popular but it is biased "toward stopping" (its mean is .5), resulting rule may be too aggressive.
- We should choose so that $(\alpha+\beta)$ is small, eg. take $\alpha=1$, but not easy to set the mean $\alpha/(\alpha+\beta)$
- (i) setting $\alpha/(\alpha+\beta) = \pi_A$ may also be somewhat biased toward stopping- unless want more cautious,
- (ii) setting $\alpha/(\alpha+\beta) = \pi_0$ may be biased toward non-stopping; may be this is the choice when investigator believe that the treatment is safe.

REVISED BAYESIAN RUKE

Suppose we choose, as prior, $\alpha=1$ and $\beta=m$ (eg. $m=32$ so that the prior mean is $\alpha/(\alpha+\beta) = \pi_0 = .03$); the revised rule is:

$$P(\pi_*) = \Pr(\pi > \pi_*)$$

$$= 1 - \int_0^{\pi_*} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+e)\Gamma(\beta+n-e)} y^{e+\alpha-1}(1-y)^{n-e+\beta-1}\,dy$$

$$= \sum_{i=0}^{e} b[i; n+m, \pi_*]$$

The trial is stop when $P(\pi_A)$ is large, say exceeding 80% or 90%; π_A being the ceiling side-effect's rate.

EXAMPLE

If we choose m=32 so that the prior mean is $\alpha/(\alpha+\beta) = \pi_0 = .03$), then:

$$.80 = \binom{n(1)+32}{0} \pi_A^0 (1-\pi_A)^{n(1)+32-0} + \binom{n(1)+32}{1} \pi_A^1 (1-\pi_A)^{n(1)+32-1}$$

$$.80 = (1-\pi_A)^{n(1)+32} + [n(1)+32]\pi_A (1-\pi_A)^{n(1)+31}$$

$n(1)$ is negative, no stopping - just as in the test based rule

This choice would result in a rule which is even more conservative the test-based one.

EXAMPLE

If we choose m=6 so that the prior mean is $\alpha/(\alpha+\beta) = \pi_A = .15$), then:

$$.80 = \binom{n(1)+6}{0}\pi_A^0(1-\pi_A)^{n(1)+6-0} + \binom{n(1)+6}{1}\pi_A^1(1-\pi_A)^{n(1)+6-1}$$

$$.80 = (1-\pi_A)^{n(1)+6} + [n(1)+6]\pi_A(1-\pi_A)^{n(1)+5}$$

n(1) is still negative, no stopping

I believe that this choice would result in a rule which is closer to the hypothesis test-based one.

After a rule is formed, including the Bayesian rule, we can always calculate its type I error rate and check to see if it is over aggressive.

$$1 - \Pr(\mathrm{N}; \pi_0, \mathit{Rule})$$

EXERCISE

Suppose we are conducting a small phase II trial with N=25 patients. We wish to form a sequential stopping rule with these two parameters: $\pi_0 = .05$ and $\pi_A = .20$

T7.1 For a rule by applying the SPRT and calculate its power and its type I error rate.

T7.2 For a rule by applying Mehta and Cain's Bayesian rule and calculate type I error rate.

T7.3 For a Bayesian rule by choosing $\alpha = 1$ and $\beta = m$ so that $\alpha/(\alpha+\beta) = \pi_0 = .05$; Calculate type I error rate.

T7.4 For a Bayesian rule by choosing $\alpha = 1$ and $\beta = m$ so that $\alpha/(\alpha+\beta) = \pi_A = .20$; Calculate type I error rate.

REFERENCES

- Armitage P. 91975). Sequential Medical Trials. New York: Wiley
- Metah CR and Cain KC (1984). Charts for the early stopping of pilots studies. J Clinical Oncology 2: 676-692.
- Goldman AI (1987). Issues in designing sequential stopping rules for monitoring side effects in clinical trials. Controlled Clinical Trials 8: 327-337.
- Thall PF, Simon RM, and EH Estey (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J Clinical Oncology 14: 296-303
- Goldman AI and PJ Hannan (2001). Optimal continuous sequential boundaries for monitoring toxicity in clinical trials: a restricted search algorithm. Statistics in Medicine 20: 1575-1589.